JOURNAL
OF THE
WASHINGTON ACADEMY OF SCIENCES
VOL. 31 MARCH 15, 1941 NO. 3

STATISTICS.—*Some thoughts on statistical inference.*[1] W. EDWARDS DEMING, Bureau of the Census.

Definition of a statistician.—In the development of the scientific method, it is usually assumed that all observations give the same result; e.g., that $F = ma$ exactly, always. The statistical method is the scientific method, modified—that is, brought up to date—to take account of the fact that all nature is subject to variations. The chief duty of a statistician is to study these variations and to design experiments so that they may provide the maximum knowledge for purposes of prediction; another is to compile data with the same object in view; and still a third function is to help bring about changes in the sources of the data. You can go back and substitute the word scientist for statistician, and have a good definition of a scientist.

Anyone who is interested in getting the most out of an experiment, and presenting the data in such form that they can be used for meaningful predictions, is something of a statistician. A qualified statistician, however, in addition to being accomplished in some branch of science, natural or social, must also be trained in probability and the mathematics of distribution theory. He must get in and work with the scientist and be one. Some statisticians, I suppose, are better than others, but the best statisticians are the best scientists. The statistician and the scientist have the same ultimate object in view, and they must work together under the same rules. The statistician has no special license. Often his special training in distribution theory is not so much help to a statistician as his training in other topics of science.

Every interpretation of data involves a prediction.—Scientific data have no meaning until they are interpreted, and there can be no interpretation except in a predictive sense. There is no such thing as scientific data merely as facts. The interpretation can not be separated from the prediction.

[1] From a discourse delivered at the National Bureau of Standards on November 22, 1940. In substance the same material was presented at the Secchi Academy of Georgetown University on February 7, 1940. Received January 27, 1941.

The idea of presenting experimental results as original data is familiar to all of us. However, presentation as a prediction may not be so familiar; in fact some scientists and engineers may prefer to think of only two ways of presenting the results of experimental work, namely, as original data and as an interpretation. Closer examination reveals, however, that every meaningful interpretation involves a prediction. (Shewhart,[2] Ch. 3.)

Moreover,

. . . there is no knowledge of external reality without the anticipation of future experience. . . . There is no knowledge without interpretation . . . what the concept denotes has always some temporal spread. (Lewis,[3] p. 195.)

Such simple concepts as blue and round, for example, embrace not an immediate quale, but some stable pattern of relations . . . Feeling the roundness of a marble as we roll it between thumb and fingers, or measuring a house, is again a temporally extended and ordered relation of apprehended qualia . . . The ascription of a substantive or an adjective is the hypothesis of some sequence in possible experience, or a multiplicity of such experiences. (Lewis,[3] pp. 129 and 132.)

Every set of data, for generalizations and conclusions, is but a sample, and a sample of the past.—One may take data just for an inventory. There are times when it is highly important to have an inventory, as for purposes of taxation. Otherwise one takes data with the object of saying something about future data that will arise from the same cause systems, or of doing something about the source of the data so that future data will differ in a certain way from past data. You design and carry out experiments on the specific heat of steam, the fatigue of metals, the pitting of metal pipe, the disintegration of leather, not just to learn something about what the specific heat of steam was, or about those particular specimens, or the particular batches whence they were drawn, all of which were made in the past, but rather, to say what the specific heat of steam is going to be next month, or to help somebody make better or more uniform steel, leather, or pipe, in the future.

Usually you do not run experiments on all the materials and articles of a particular batch that were made in the past; you do not need to; you experiment on only part of them. You draw a sample. But even if you ran experiments on an entire batch, i.e., took a 100 percent sample of last month's production, you still would have only a sample of what would have been produced by the same machines or exactly similar machines, reoperated under the same essential conditions. You are obliged to experiment on materials or articles that were made in the past, with the object of drawing inferences about some that are to be made in the future.

[2] Walter A. Shewhart. *Statistical method from the viewpoint of quality control.* (The Graduate School, Department of Agriculture, Washington, 1939.)

[3] C. I. Lewis. *Mind and the world-order.* (Scribners, New York, 1929.)

The collection of data, whether by a physicist in the laboratory, by a government census of population, agriculture, or unemployment, by a department store on its sales and complaints and quality of goods placed on the shelves, by a manufacturer who keeps records of the quality of his products, or by any other person, is for the ultimate purpose of taking some sort of action, or making recommendation for action. Even a classroom quiz is no exception; it is not given (we hope) just for the nuisance of it, but to enable the teacher to say, on the basis of past evidence, whether certain pupils will be able to go ahead in the future into more advanced work, or to hold a job. The teacher, if he is scientifically minded, will have still another object in view, namely, to improve his own teaching; by seeing how well or how poorly his pupils have picked up certain ideas that he has tried to instill into them in the past, he can judge his own work to see where he needs to adopt different methods of teaching in the future.

In taking readings with a galvanometer you can always conceivably take one more reading, but in actual practice you are satisfied with a finite number. From these readings taken in the past, you make statements about what someone else will find when he takes readings in the future. Whatever special studies are pursued by statisticians, the chief object of learning them is to acquire facility in making predictions from data, and in presenting data in such form that others can do the same.

What I am saying was voiced more succinctly by Fry of the Bell Telephone Laboratories, at the University of Pennsylvania Bicentennial Celebration:

> The statistical method is used for saying something about data that we are about to take, not what we have already taken.[4]

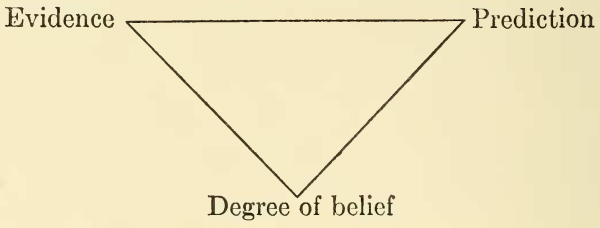
A word on sampling and the census.—So far as scientific generalizations and predictions are concerned, the distinction between a sample and a complete count (a perfect census, tests run on all of last month's production, all the readings that you might have taken with your galvanometer but did not), is only one of degree. A complete count of last month's production is only a bigger sample than part of it. Both are samples of what might have resulted, and the kind of results that are to be expected in the future from the same underlying cause system. In population studies, the births, deaths, vocations, migrations, and educational attainments of a population are changed and directed by a myriad of chance causes, superimposed on certain underlying

[4] Thornton C. Fry. September 19, 1940. The quotation given may not be verbatim.

social and economic cause systems. A census, even if perfect, is subject to the variations of chance because it describes only one of the many possible populations that might have been found as the result of this combination of chance causes and main causes. Any generalizations (i.e., predictions) must recognize the fact that some other population might have resulted, and in fact must be expected to arise in the future, from the same underlying causes.[5]

How big a sample? The three components of knowledge.—Seeing, then, that we must make our predictions about the future from samples of the past, the question arises: How big must these samples be? How much data do we need? One distinguishing characteristic of a good scientist, I should say, is that he knows good data and knows when he has enough of it.

But how much is enough? 10, 20, 50, 100, or 1,000 readings? The answer is not a simple yes or no. It is tied up with the idea that a prediction, if it is to convey knowledge, must be based on evidence. The degree of belief in any prediction is closely linked with the prediction and the evidence. Shewhart (*op. cit.*, p. 86) exhibits the triangular relation shown below, linking the three components of knowledge:



THE THREE COMPONENTS OF KNOWLEDGE

On the basis of certain evidence, you would make certain predictions, and in so doing, convey a certain degree of belief. A prediction is expressed in terms of data that one would expect to get if he were to perform certain experiments in the future. A prediction without any supporting evidence conveys no degree of belief. Thus, if I say it is going to rain day after tomorrow, I have made a prediction, but created no degree of belief, because you have no evidence, since I have no standing as a weather prophet. You would likely not carry your umbrella or cancel your trip on the basis of my prediction. The results of experimental work are usually summarized in terms of predictions and evidence. Shewhart gives a rule for the presentation of data, stat-

[5] This topic is pursued in more detail in a paper by W. EDWARDS DEMING and FREDERICK F. STEPHAN in the Journal of the American Statistical Association, March 1941.

ing that original data, if summarized, should be summarized in such a way that the evidence in the original data is preserved for all the predictions that are thought to be useful. Judgment is of course required in regard to just which predictions are to be assumed useful, but this is where scientific judgment must be exercised.

The question before us is how much data does one need? We are now ready to get back to it and look for an answer. I should say that one needs enough data—i.e., evidence—to provide some substantial degree of belief in whatever predictions he chooses to make or expects others to make. A physicist usually does not commence to record data for publication or calibration until his apparatus has been perfected to a state satisfactory to him; and he stops when further data, in his judgment, would create no greater degree of belief in any prediction that he may wish to make from that experiment.

Statistical control.—In the state of statistical control or the state of randomness, the data from an experiment, or the measurements on a product manufactured, display the characteristics of statistical stability. They behave as if they were samples being drawn at random from a stable universe. The ideal stable universe is a bowl of physically similar numbered chips. When they are thoroughly stirred and drawn blindfolded with replacement, the resulting sequence of numbers is the result of a random operation. (The necessity for extreme care in attempting to carry out any random operation, even in so simple an experiment as drawing numbers from a bowl, can hardly be overemphasized.) In experimental work, and in manufacturing, one can not form a judgment in regard to the attainment of statistical control until his experiment or process has been continued long enough to be subjected to Shewhart's Criterion I, which requires at least 100 readings.[6]

In the state of statistical control, the distribution theories of mathematical statistics apply, and it is possible to make a valid prediction concerning the next hundred or thousand observations; it is possible, for instance, to draw a pair of limits (control limits) such that whenever a future observation falls outside these limits, it will be worth while to look for an assignable cause of variation in the process.

The state of statistical control is the goal of all economic manufacture of materials. It is not usually the goal in experimental work in pure science; but this is only an illustration of the fact that some

[6] This criterion for randomness is discussed by Shewhart in his *Economic control of quality* (Van Nostrand, 1931), pp. 304–318. For a description of a "normal" bowl, and the results of 4,000 drawings therefrom, see Shewhart's *Economic control of manufactured product* (Van Nostrand, 1931), table 22 and Appendix II.

requirements of industry are more exacting than the requirements of pure science. Usually an experimental physicist or chemist is satisfied to eliminate trends and erratic disturbances in his experiment to a point wherein he feels confident that he can draw a pair of limits that include any future observations that he might make by that method. He may continue making adjustments until he can set these limits narrower than has been possible in any previous similar experiment. If, with still further adjustments, his experiment were brought to a state of statistical control, still narrower limits could be set with even greater confidence. However, we shall find that the course adopted by the pure scientist is very often justifiable from the standpoint of accuracy, for the reason that any limits that he might draw on the basis of one experiment refer only to what may be expected from that particular method, but they do not by themselves give any indication of systematic errors nor of what may be expected from some other method of measurement.

Accuracy and precision.—Scientists have for long toyed with the hope of finding some logical method of inferring from a given set of data what the accuracy of those data may be. The idea is an exciting one, but it is a vain hope; the data of a single experiment, or even of a number of experiments, do not by themselves provide all the evidence that is needed for stating the true value of the thing being measured. For to say something about the true value, one must predict what will be the result of all other methods of measurement, not only those methods that have been tried out, but also all those that are yet to be devised. He must also be able to explain any discrepancies between different methods. More methods and more data (good data) add new evidence to our knowledge, but of knowledge there is no end. The concept of true value arises not from any highly consistent results arising from one experiment, but from consistent results from many different kinds of experiments. Think of the different ways of measuring e/m. When two or three of them had been discovered, and found to give consistent results, there was reason to begin thinking that something was known about e/m. But new data and new methods can always upset predictions, and such has been the history of physics. "Knowing begins and ends in experience; but it does not end in the experience in which it began."[7]

The objectivity of being able to make a valid prediction of the limits within which the future data of a single method of measurement will fall is in contrast with the subjectivity of assigning limits within

[7] C. I. Lewis. *Experience and meaning,* Philos. Rev. **13**: 134. 1934.

which the data of that and all other methods will fall. This contrast is expressed by the distinction between the words precision and accuracy. The limits set by a single experiment give a measure of the precision of that experiment, or of that method of measurement, but they give no objective criterion of the accuracy of the result, because they refer to the one experiment only, and not to all the other methods that are or have been devised.

In order to convey knowledge, the Δx in $x \pm \Delta x$ must have an operationally verifiable meaning in the form of a prediction. This will be one sort of a prediction in a statement regarding precision, but a very different one in a statement regarding accuracy.

To see how these remarks apply in practice, let us think of a certificate issued by the National Bureau of Standards on a precision standard.

UNITED STATES DEPARTMENT OF COMMERCE
WASHINGTON
NATIONAL BUREAU OF STANDARDS

Certificate
for
Standard Resistor

Manufactured by _____, Serial No. _____

Submitted by

The _____ _____ Company

of _____

The above-described resistor was found in September 1940, at a temperature of 25°C, to have a resistance of

9.9999 International Ohms.

The value given is correct within 0.005 percent. This statement of accuracy takes into consideration the uncertainty in the realization of the International Ohm from its definition and allows for normal changes of resistance with time.

LYMAN J. BRIGGS, *Director*

Let us try to see the element of prediction in this certificate. Perhaps we shall agree that a certificate is a prediction. When a piece of apparatus is of such poor quality that you dare not risk a prediction on it, you do not issue a certificate, but may instead issue a report. In issuing a certificate you risk making a prediction regarding the future behavior of a piece of apparatus that was sent in for test; in a report you merely record a bit of history—how it behaved, in terms of your own standards—while it was here. You leave the risk of prediction to the owner of the apparatus.

In the testing of precision standard resistors at the National Bureau

of Standards, the measurements can ordinarily be duplicated within a range of 1 part in 100,000 to 1 part in 10,000,000, depending partially on the magnitude of the resistor, but more particularly on its quality. In the example given above, the measurements can be duplicated from day to day with variations of not more than one unit in the fifth decimal place, and the resistance is therefore determined in terms of the N.B.S. Ohm to within this magnitude. These are statements of precision and are objective. As for accuracy—comparison with the International Ohm, and behavior after being shipped back to its owner—you are obliged to depend on intuition. You allow a wide factor of safety; your certificate risks a prediction that is 50 times as wide as the latitude of reproducibility of your measurements. You did not make a direct comparison with the International Ohm, and you did not run a test on the change of resistance of this particular resistor with time, and certainly not after it was sent back to its owner. But you have had many years' experience with similar resistors, and you are led by intuition to make statements (predictions) regarding the accuracy of this particular one. You feel safe in predicting its behavior. For resistors of better quality, you would name narrower limits, and for one of particularly good quality, you might even make a prediction regarding the variation of its resistance with temperature.

In a certificate you are not talking about your own apparatus: you are talking about the apparatus that was sent in for test.

Intuition may at times be very helpful, but intuition, like the conscience, must be trained. The distribution theory of statistics should be a part of this training. The rest of us may well be satisfied to pin our faith on the intuition of an expert. And perhaps our own intuition helps us to distinguish between experts and others whose intuition is not so reliable.

Pure distribution theory, by itself, is nigh helpless until the state of statistical control is attained and proved. Since statistical control seldom exists in experimental work, the interpretation of scientific data remains, for the most part, a matter of cooperation between the statistician and the scientist, each assisting the other in the process of adjusting the apparatus, and finally in making predictions from the results. The peculiar training of the statistician enables him to help the physicist or engineer to weed out assignable causes of variation and to attain uniformity; in fact, as I said, one of the chief duties of a statistician is to help bring about desirable changes in the source of the data that he takes. His services are especially useful in industry,